Unifying Divergence Minimization and Statistical Inference via Convex Duality

Yasemin Altun[1] and Alex Smola[2]

[1] Toyota Technological Institute at Chicago, Chicago, IL 60637, USA**
altun@tti-c.org
[2] National ICT Australia, North Road, Canberra 0200 ACT, Australia
alex.smola@nicta.com.au

Abstract. In this paper we unify divergence minimization and statistical inference by means of convex duality. In the process of doing so, we prove that the dual of approximate maximum entropy estimation is maximum a posteriori estimation. Moreover, our treatment leads to stability and convergence bounds for many statistical learning problems. Finally, we show how an algorithm by Zhang can be used to solve this class of optimization problems efficiently.

1 Introduction

It has become part of machine learning folklore that maximum entropy estimation and maximum likelihood are convex duals to each other. This raises the question whether maximum a posteriori estimates have similar counterparts and whether such estimates can be obtained efficiently.

Recently Dudik et al. [9] showed that a certain form of regularized maximum entropy density estimation corresponds to ℓ_1 regularization in the dual problem. This is our starting point to develop a theory of regularized divergence minimization which aims to unify a collection of related approaches. By means of convex duality we are able to give a common treatment to

- The regularized LMS minimization methods of Arsenin and Tikhonov [21], and of Morozov [14]. There the problem of minimizing

$$\|x\|_2^2 \text{ subject to } \|Ax - b\|_2^2 \leq \epsilon$$

  is studied as a means of improving the stability of the problem $Ax = b$.
- Ruderman and Bialek [18] study a related problem where instead of a quadratic penalty on x the following objective function is minimize

$$-H(x) \text{ subject to } \|Ax - b\|_2^2 \leq \epsilon$$

  In other words, the problem of solving $Ax = b$ is stabilized by finding the maximum entropy distribution which satisfies the constraint.

** Parts of this work were done when the author was visiting National ICT Australia.

– The density estimation problem of [9] can be viewed as one of solving a variant of the above, namely that of minimizing

$$-H(x) \text{ subject to } \|Ax - b\|_\infty \leq \epsilon$$

where the constraint encode deviations of the measured values of some *moments* or *features* and their expected values.

The problem we study can be abstractly stated as the regularized inverse problem

$$\underset{x \in \mathcal{X}}{\text{minimize}}\, f(x) \text{ subject to } \|Ax - b\|_{\mathcal{B}} \leq \epsilon.$$

where $\mathcal{X}$ and $\mathcal{B}$ are Banach spaces. We start by establishing a general framework of duality to solve this problem using a convex analysis tool, namely Fenchel's duality. This theory is especially useful in the most general form of our problem, where $\mathcal{X}$ and $\mathcal{B}$ are infinite dimensional, since in this case Lagrangian techniques are problematic due to differentiability issues. We apply this framework to a generalized notion of regularized divergence minimization, since a large subset of statistical learning literature can be analyzed within this class of problems.

By studying convex duality of two important classes of divergences, namely Csiszár and Bregman divergences, we show that maximum a posteriori estimation is the convex dual of approximate maximum entropy estimation. Various statistical inference methods, such as boosting, logistic regression, Gaussian Processes and others become instances of our framework, by using different entropy functions and regularization methods. Following these lines, we not only give a common treatment to these methods, but also provide directions to develop new inference techniques by investigating different entropy-regularization combinations. For example, working in Banach spaces, we can perform different regularizations on subsets of basis functions, which is useful in problems like structured learning where there are several distinctly different sets of basis functions.

From a regularization point of view, our approach provides a natural interpretation to the regularization coefficient ϵ, which corresponds to the approximation parameter in the primal problem. Studying the concentration of empirical means, we show that a good value of ϵ is proportional to $O(1/\sqrt{m})$ where m is the sample size. Noting that ϵ is generally chosen by cross validation techniques in practice, we believe our framework gives us an enhanced interpretation of regularized optimization problems. We also provide unified bounds on the performance of the estimate wrt loss on empirical estimates as well as the loss on true statistics, which apply to arbitrary linear classes and divergences. Finally, we show that a single algorithm can efficiently optimize this large class of optimization problems with good convergence rates.

Related work There is a large literature on analyzing various loss functions on exponential families as the convex dual of relative entropy minimization via equality constraints of the form $Ax = b$. For example, Lafferty [12] analyze logistic regression and exponential loss as a special case of Bregman divergence minimization and propose a family of sequential update algorithms. Similar treat-

ments are given in [11, 8]. One common property of these studies is that they investigate exact divergence minimization.

Previous work on approximate divergence minimization focused on minimizing KL divergence such that its convex dual is penalized by ℓ_1 and ℓ_2 norm terms, eg. [7]. [9] show that approximate KL divergence minimization wrt. $\|Ax - b\|_\infty \leq \epsilon$ has the convex dual of ℓ_1 norm regularized maximum likelihood. Recently [10] produced similar results for ℓ_p norm regularization.

In this paper, we improve over previous work by generalizing to a family of divergence functions with inequality constraints in Banach spaces. Our unified treatment of various entropy measures (including Csiszár and Amari divergences) and various normed spaces allows us to produce the cited work as special cases and to define more sophisticated regularizations via Banach space norms. Finally we provide risk bounds for the estimates.

2 Fenchel Duality

We now give a formal definition of the class of inverse problems we solve. Denote by $\mathcal{X}$ and $\mathcal{B}$ Banach spaces and let $A : \mathcal{X} \to \mathcal{B}$ be a bounded linear operator between those two spaces. Here A corresponds to an "observation operator", e.g. mapping distributions into a set of moments, marginals, etc. Moreover, let $b \in \mathcal{B}$ be the target of the estimation problem. Finally, denote by $f : \mathcal{X} \to \mathbb{R}$ and $g : \mathcal{B} \to \mathbb{R}$ convex functions and let $\epsilon \geq 0$.

Problem 1 (Regularized inverse). *Our goal is to find $x \in \mathcal{X}$ which solves the following convex optimization problem:*

$$\underset{x \in \mathcal{X}}{\text{minimize}}\, f(x) \text{ subject to } \|Ax - b\|_{\mathcal{B}} \leq \epsilon.$$

Example 2 (Density estimation). *Assume that x is a density, f is the negative Shannon-Boltzmann entropy, b contains the observed values of some* moments *or* features*, A is the expectation operator of those features wrt. the density x and the Banach space $\mathcal{B}$ is ℓ_p.*

We shall see in Section 3.2 that the dual to Example 3 is a maximum a posteriori estimation problem.

In cases where $\mathcal{B}$ and $\mathcal{X}$ are finite dimensional the problem is easily solved by calculating the corresponding Lagrangian, setting its derivative to 0 and solveing for x. In the infinite dimensional case, more careful analysis is required to ensure continuity and differentiability. Convex analysis provides a powerful machinery, namely Fenchel's conjugate duality theorem, to study this problem by formulating the primal-dual space relations of convex optimization problems in our general setting. We need the following definition:

Definition 3 (Convex conjugate). *Denote by $\mathcal{X}$ a Banach space and let $\mathcal{X}^*$ be its dual. The convex conjugate or the Legendre-Fenchel transformation of a function $f : \mathcal{X} \to \mathbb{R}$ is $f^* : \mathcal{X}^* \to \mathbb{R}$ where f^* is defined as*

$$f^*(x^*) = \sup_{x \in \mathcal{X}} \left\{ \langle x, x^* \rangle - f(x) \right\}.$$

We present Fenchel's theorem where the primal problem is of the form $f(x) + g(Ax)$. Problem 2 becomes an instance of the latter for suitably defined g.

Theorem 4 (Fenchel Duality [3, Th. 4.4.3]). *Let $g : \mathcal{B} \to \mathbb{R}$ be a convex function on $\mathcal{B}$ and other variables as above. Define t and d as follows:*

$$t = \inf_{x \in \mathcal{X}} \{f(x) + g(Ax)\} \text{ and } d = \sup_{x^* \in \mathcal{B}^*} \{-f^*(A^*x^*) - g^*(-x^*)\}.$$

Assume that f, g and A satisfy one of the following constraint qualifications:
a) $0 \in \operatorname{core}(\operatorname{dom} g - A \operatorname{dom} f)$ *and both f and g are left side continuous (lsc)*
b) $A \operatorname{dom} f \cap \operatorname{cont} g \neq \emptyset$
Here $s \in \operatorname{core}(S)$ if $\bigcup_{\lambda > 0} \lambda(S - s) \subseteq \mathcal{X}$ where $\mathcal{X}$ is a Banach space and $S \subseteq \mathcal{X}$. In this case $t = d$, where the dual solution d is attainable if it is finite.

We now apply Fenchel's duality theorem to convex constraint optimization problems, such as Problem 2, since the dual problem is easier to solve in certain cases.

Lemma 5 (Fenchel duality with constraints). *In addition to the assumptions of Theorem 5, let $b \in \mathcal{B}$ and $\epsilon \geq 0$. Define t and d as follows:*

$$t = \inf_{x \in \mathcal{X}} \{f(x) \text{ subject to } \|Ax - b\|_{\mathcal{B}} \leq \epsilon\}$$
$$\text{and } d = \sup_{x^* \in \mathcal{B}^*} \{-f^*(A^*x^*) + \langle b, x^* \rangle - \epsilon \|x^*\|_{\mathcal{B}^*}\}$$

Suppose f is lower semi-continuous and that for $B := \left\{\bar{b} \in \mathcal{B} \text{ with } \left\|\bar{b}\right\| \leq 1\right\}$ the following constraint qualification holds:

$$\operatorname{core}(A \operatorname{dom} f) \cap (b + \epsilon \operatorname{int}(B)) \neq \emptyset. \qquad (CQ)$$

In this case $t = d$ with dual attainment.

Proof Define g in Theorem 5 as the characteristic function on $\epsilon B + b$, i.e.

$$g(\bar{b}) = \chi_{\epsilon B + b}(\bar{b}) = \begin{cases} 0 & \text{if } \bar{b} \in \epsilon B + b \\ \infty & \text{otherwise} \end{cases} \qquad (1)$$

The convex conjugate of g is given by

$$\begin{aligned} g^*(x^*) &= \sup_{\bar{b}} \left\{\left\langle \bar{b}, x^* \right\rangle \text{ subject to } \bar{b} - b \in \epsilon B\right\} \\ &= -\langle x^*, b \rangle + \epsilon \sup_{\bar{b}} \left\{\left\langle \bar{b}, x^* \right\rangle \text{ subject to } \bar{b} \in B\right\} = \epsilon \|x^*\|_{\mathcal{B}^*} - \langle x^*, b \rangle \end{aligned}$$

Theorem 5 and the relation $\operatorname{core}(B) = \operatorname{int}(B)$ prove the lemma. ■

The constraint qualification (CQ) ensures the non-emptiness of the sub-differential. $\epsilon = 0$ leads to equality constraints $Ax = b$, for which CQ requires b to be an element of $\operatorname{core}(A \operatorname{dom} f)$. If the equality constraints are not feasible $b \notin \operatorname{core}(A \operatorname{dom} f)$, which can be the case in real problems, the solution diverges.

Such problems may be rendered feasible by relaxing the constraints ($\epsilon > 0$), which corresponds to expanding the search space by defining an ϵ ball around b and searching for a point in the intersection of this ball and $\operatorname{core}(A \operatorname{dom} f)$. In the convex dual problem, this relaxation is penalized with the norm of the dual parameters scaling linearly with the relaxation parameter ϵ.

In practice it is difficult to check whether (CQ) holds. One solution is to solve the dual optimization problem and infer that the condition holds if the solution does not diverge. To assure a finite solution, we restrict the function class such that f^* is Lipschitz and to perturb the regularization slightly by taking its k^{th} power, resulting in a Lipschitz continuous optimization. For instance Support Vector Machines perform this type of adjustment to ensure feasibility [4].

Lemma 6. *Denote by $\mathcal{X}$ a Banach space, let $b \in \mathcal{X}^*$ and let $k > 1$. Assume that $f(Ax)$ is convex and Lipschitz continuous in x with Lipschitz constant C. Then*

$$\inf_{x \in \mathcal{X}} \left\{ f(Ax) - \langle b, x \rangle + \epsilon \|x\|^k \right\} \tag{2}$$

does not diverge and the norm of x is bounded by $\|x\|_{\mathcal{X}} \leq \left[\left(\|b\|_{\mathcal{X}^*} + C \right) / k\epsilon \right]^{\frac{1}{k-1}}$.

Proof [sketch] Note that the overall Lipschitz constant of the objective function (except for the norm) is bounded by $\|b\|_{\mathcal{X}^*} + C$. The objective function cannot increase further if the slope due to the norm is larger than what the Lipschitz constant admits. Solving for $\epsilon k \|x\|_{\mathcal{X}}^{k-1} = \|b\|_{\mathcal{X}^*} + C$ proves the claim. ■

3 Divergence Minimization and Convex Duality

Now that we established a rather general framework of duality for regularized inverse problems, we consider applications to problems in statistics. For the remainder of the section x will either be a density or a conditional density over the domain $\mathcal{T}$. For this reason we will use p instead of x to denote the variable of the optimization problem.

Denote by $\psi : \mathcal{T} \to \mathcal{B}$ feature functions and let $A : \mathcal{X} \to \mathcal{B}$ be the expectation operator of the feature map with respect to p. In other words, $Ap := \mathbf{E}_{t \sim p} [\psi(t)]$. With some abuse of notation we will use the shorthand $\mathbf{E}_p[\psi]$ whenever convenient. Finally denote by $\tilde{\psi} = b$ the observed value of the features $\psi(t)$, which are derived, e.g. via $b = m^{-1} \sum_{i=1}^{m} \psi(t_i)$ for $t_i \in S$, the sample of size m.

This setting allows us to study various statistical learning methods within convex duality framework. It is well-known that the dual of maximum (Shannon) entropy (commonly called MaxEnt) is maximum likelihood (ML) estimation. One of the corollaries, which follows immediately from the more general result in Lemma 12 is that the dual of *approximate* maximum entropy is maximum *a posteriori* estimation (MAP).

Theorem 7. *Assume that f is the negative Shannon entropy, that is $f(p) := -H(p) = \int_{\mathcal{T}} \log p(t) dp(t)$. Under the above conditions we have*

$$\min_p -H(p) \text{ subject to } \left\| \mathbf{E}_p[\psi] - \tilde{\psi} \right\| \leq \epsilon \text{ and } \int_{\mathcal{T}} dp(t) = 1 \quad (3)$$

$$= \max_\phi \left\langle \phi, \tilde{\psi} \right\rangle - \log \int_{\mathcal{T}} \exp(\langle \phi, \psi(t) \rangle) dt - \epsilon \|\phi\| + e^{-1}$$

Equivalently ϕ maximizes $\Pr(S|\phi) \Pr(\phi)$ and $\Pr(\phi) \propto \exp(-\epsilon \|\phi\|)$.

In order to provide a common treatment to various statistical inference techniques, as well as give insight into the development of new ones, we study two important classes of divergence functions, Csiszár's divergences and Bregman divergences. Csiszár divergence, which includes Amari's α divergences as special cases, gives the asymmetric distance between two infinite-dimensional density functions induced by a manifold. Bregman divergences are commonly defined over distributions over a finite domain. The two classes of divergences intersect at the KL divergence. To avoid technical problems we assume that the constraint qualifications are satisfied (e.g. via Lemma 7).

3.1 Csiszár Divergences

Definition 8. *Denote by $h : \mathbb{R} \to \mathbb{R}$ a convex lsc function and let p, q be two distributions on $\mathcal{T}$. Then the Csiszár divergence is given by*

$$f_h(q, p) = \int q(t) h\left(\frac{p(t)}{q(t)}\right) dt. \quad (4)$$

Different choices for h lead to different divergence measures. For instance $h(\xi) = \xi \log \xi$ yields the Kullback-Leibler divergence. Commonly, q is fixed and optimization is performed with respect to p, which we denote by $f_{h,q}(p)$. Since $f_{h,q}(p)$ is convex and expectation is a linear operator, we can apply Lemma 6 to obtain the convex conjugate of Csiszár's divergence optimization:

Lemma 9 (Duality of Csiszár Divergence). *Assume that the conditions of Lemma 6 hold. Moreover let f be defined as a Csiszár divergence. Then*

$$\min_p \left\{ f_{h,q}(p) | \|E_p[\psi] - \tilde{\psi}]\|_{\mathcal{B}} \leq \epsilon \right\} = \max_\phi \left\{ -f^*_{h,q}(\langle \phi, \psi(.) \rangle) + \left\langle \phi, \tilde{\psi} \right\rangle - \epsilon \|\phi\|_{\mathcal{B}^*} \right\}.$$

Moreover the solutions $\hat{p}$ and $\hat{\phi}$ are connected by $\hat{p}(t) = q(t)(h^*)'\left(\left\langle \psi(t), \hat{\phi} \right\rangle\right)$.

Proof The adjoint of the linear operator A is given by $\langle Ax, \phi \rangle = \langle A^* \phi, x \rangle$. Letting A be the expectation wrt p, we have $\left\langle \int_{\mathcal{T}} p(t)\psi(t), \phi \right\rangle = \int_{\mathcal{T}} p(t) \langle \psi(t), \phi \rangle dt = (A^*\phi)(p)$ for $A^*\phi = \langle \phi, \psi(.) \rangle$. Next note that $f^*(\langle \phi, \psi(\cdot) \rangle) = \int_{\mathcal{T}} q(t) h^*(\langle \phi, \psi(t) \rangle) dt$. Plugging this into Lemma 6, we obtain the first claim.

Using attainability of the solution it follows that there exist $\hat{p}$ and $\hat{\phi}$ which solve the corresponding optimization problems. Equating both sides we have

$$\int_{\mathcal{T}} q(t)h\Big(\tfrac{\hat{p}(t)}{q(t)}\Big)dt = -f^*(\langle\phi,\psi(.)\rangle) + \left\langle\tilde{\psi},\hat{\phi}\right\rangle - \epsilon\|\hat{\phi}\|_{B^*} = -f^*(\langle\phi,\psi(.)\rangle) + \left\langle\hat{\phi},E_{\hat{p}}[\psi]\right\rangle.$$

Here the last equality follows from the definition of the constraints (see the proof of Lemma 6). Taking the derivative at the solution $\hat{p}$ (due to constraint qualification) and noticing that the derivative of the first term on the RHS, we get $h'\left(\frac{\hat{p}}{q}\right) = \left\langle\hat{\phi},\psi\right\rangle$. Using the relation $(h')^{-1} = (h^*)'$ completes the proof. ■

Since we are dealing with probability distributions, it is convenient to add the constraint $\int_{\mathcal{T}} dp(t) = 1$. We have the following corollary.

Corollary 10 (Csiszár divergence and probability constraints). *Define all variables as in Lemma 10. We have*

$$\begin{aligned}&\min_p\left\{f_{h,q}(p)\ \textit{subject to}\ \left\|\mathbf{E}_p[\psi]-\tilde{\psi}\right\|_{\mathcal{B}}\leq\epsilon\ \textit{and}\ \int_{\mathcal{T}}dp(t)=1\right\}\\ =\ &\max_\phi\left\{-f^*_{h,q}\left(\langle\phi,\psi(.)\rangle-\Lambda_\phi\right)+\left\langle\phi,\tilde{\psi}\right\rangle-\Lambda_\phi-\epsilon\left\|\phi\right\|_{\mathcal{B}^*}\right\}=:-\mathcal{L}^C_{\tilde{\psi}}(\phi).\end{aligned}\tag{5}$$

Here the solution is given by $\hat{p}(t) = q(t)(h^*)'(\left\langle\psi(t),\hat{\phi}\right\rangle - \Lambda(\hat{\phi}))$ *where* $\Lambda(\hat{\phi})$ *is the partition function which ensures that* p *be a probability distribution (*$\lambda(\hat{\phi})$ *is the minimizer of (5) with respect to* λ_ϕ*).*

Proof [sketch] Define $\mathcal{P} = \{p|\int_{\mathcal{T}} dp(t) = 1\}$ and f in Lemma 6 as $f(p) = f_{h,q}(p) + \chi_{\mathcal{P}}(p)$. Then, for $\Lambda_p = \infty$ if $p \notin \mathcal{P}$, the convex conjugate of f is $f^*(p^*) = \sup_p\{\langle p,p^*\rangle - f_{h,q}(p) - \Lambda_p(\int_{\mathcal{T}} dp(t) - 1) = \Lambda_{p^*} + (f_{h,q})^*(p^* - \Lambda_{p^*})$. Performing the steps in the proof of Lemma 10 gives the result. ■

An important and well-studied special case of this duality is the minimization of KL divergence as we investigate in the next section. Note that new inference techniques can be derived using other h functions, eg. Tsallis' entropy, which is preferable over Shannon's entropy in fields as statistical mechanics.

3.2 MAP and Maximum Likelihood via KL Divergence

Defining h in (4) as $h(\xi) := \xi\ln(\xi)$ we have $h^*(\xi^*) = \exp(\xi^* - 1)$. Then Csiszár's divergence becomes the KL divergence. Applying Corollary 11 we have:

Lemma 11 (KL divergence with probability constraints). *Define all variables as in Lemma 11. We have*

$$\begin{aligned}&\min_p\left\{KL(p\|q)\ \textit{subject to}\ \left\|\mathbf{E}_p[\psi]-\tilde{\psi}\right\|_{\mathcal{B}}\leq\epsilon\ \textit{and}\ \int_{\mathcal{T}}dp(t)=1\right\}\\ =\ &\max_\phi\left\{\left\langle\phi,\tilde{\psi}\right\rangle-\log\int_{\mathcal{T}}q(t)\exp(\langle\phi,\psi(t)\rangle)dt-\epsilon\|\phi\|_{\mathcal{B}^*}+e^{-1}\right\}\end{aligned}\tag{6}$$

where the unique solution is given by $\hat{p}_{\hat{\phi}}(t) = q(t)\exp\left(\left\langle\hat{\phi},\psi(t)\right\rangle - \Lambda_{\hat{\phi}}\right)$.

Proof The dual of f is $f^*_{h,q}(x^*) = \int_{\mathcal{T}} q(t)\exp(x^*(t)-1)dt$. Hence we have the dual objective function

$$\int_{\mathcal{T}} q(t)\exp\left(\langle\phi,\psi(t)\rangle - \Lambda_\phi - 1\right)dt + \left\langle\phi,\tilde{\psi}\right\rangle - \Lambda_\phi - \epsilon\left\|\phi\right\|_{\mathcal{B}^*}$$

We can solve for optimality in Λ_ϕ which yields $\Lambda_\phi = \log\int_{\mathcal{T}} q(t)\exp\left(\langle\phi,\psi(t)\rangle\right)dt$. Substituting this into the objective function proves the claim. ■

Thus, optimizing approximate KL divergence leads to exponential families. Many well known statistical inference methods can be viewed as special cases. Let $\mathcal{P} = \{p | p \in \mathcal{X}, \int_{\mathcal{T}} dp(t) = 1\}$ and $q(t) = 1, \forall t \in \mathcal{T}$.

Example 12. *For $\epsilon = 0$, we get the well known duality between Maximum Entropy and Maximum Likelihood estimation.*

$$\min_{p\in\mathcal{P}}\left\{-H(p)\ \textit{subject to}\ \mathbf{E}_p\left[\psi\right] = \tilde{\psi}\right\} = \max_{\phi}\left\langle\phi,\tilde{\psi}\right\rangle - \log\int_{\mathcal{T}}\exp(\langle\phi,\psi(t)\rangle)dt + e^{-1}$$

Example 13. *For $\mathcal{B} = \ell_\infty$ we get the density estimation problem of [9]*

$$\begin{aligned}&\min_{p\in\mathcal{P}}\left\{-H(p)\ \textit{subject to}\ \left\|\mathbf{E}_p\left[\psi\right] - \tilde{\psi}\right\|_\infty \le \epsilon\right\}\\ =&\max_{\phi}\left\langle\phi,\tilde{\psi}\right\rangle - \log\int_{\mathcal{T}}\exp(\langle\phi,\psi\rangle\left(t\right))dt - \epsilon\|\phi\|_1 + e^{-1}\end{aligned}$$

If $\mathcal{B}$ is a reproducing kernel Hilbert space of spline functions we obtain the density estimator of [16], who use an RKHS penalty on ϕ.

The well-known overfitting behavior of ML can be explained by the constraint qualification (*CQ*) of Section 2. While it can be shown that in exponential families the constraint qualifications are satisfied [22] if we consider the closure of the marginal polytope, the solution may be on (or close to) a vertex of the marginal polytope. This can lead to large (or possibly diverging) values of ϕ. Hence, regularization by approximate moment matching is useful to ensure that such divergence does not occur.

Regularizing ML with ℓ_2 and ℓ_1 norm terms is a common practice [7], where the coefficient is determined by cross validation techniques. The analysis above provides a unified treatment of the regularization methods. But more importantly, it leads to a principled way of determining the regularization coefficient ϵ as discussed in Section 4.

Note that if $t \in \mathcal{T}$ is an input-output pair $t = (x, y)$ we could maximize the entropy of either the *joint* probability density $p(x, y)$ or the *conditional* model $p(y|x)$, which is what we really need to estimate $y|x$. If we maximize the entropy of $p(y|x)$ and $\mathcal{B}$ is a RKHS with kernel $k(t, t') := \langle\psi(t), \psi(t')\rangle$ we obtain a range of conditional estimation methods:

- For $\psi(t) = y\psi_x(x)$ and $y \in \{\pm 1\}$, we obtain binary Gaussian Process classification [15].

- For $\psi(t) = (y, y^2)\psi_x(x)$, we obtain the heteroscedastic GP regression estimates of [13].
- For decomposing $\psi(t)$, we obtain various graphical models and conditional random fields as described in [1].
- For $\psi(t) = y\psi_x(x)$ and ℓ_∞ spaces, we obtain as its dual ℓ_1 regularization typically used in sparse estimation methods.

The obvious advantage of using convex duality in Banach spaces is that it provides a unified approach (including bounds) for different regularization/relaxation schemes as listed above. More importantly, this generality provides flexibility for more complex choices of regularization. For instance, we could define different regularizations for features that possess different characteristics.

3.3 Bregman Divergence

The Bregman divergence between two distributions p and q for a convex function h acting on the space of probabilities is given by

$$\triangle_h(p, q) = h(p) - h(q) - \langle (p - q), \nabla_q h(q) \rangle . \tag{7}$$

Note $\triangle_h(p, q)$ is convex in p. Applying Fenchel's duality theory, we have

Corollary 14. Duality of Bregman Divergence *Assume that the conditions of Lemma 6 hold. Moreover let f be defined as a Bregman divergence. Then*

$$\begin{aligned} &\min_p \left\{ \triangle_h(p, q) \text{ subject to } \left\| \mathbf{E}_p[\psi] - \tilde{\psi} \right\|_{\mathcal{B}} \leq \epsilon \right\} \\ &= \max_\phi \left\{ -h^* \left(\langle \phi - \phi_q, \psi \rangle \right) + \left\langle \phi, \tilde{\psi} \right\rangle - \epsilon \left\| \phi \right\|_{\mathcal{B}^*} \right\} =: -\mathfrak{L}^{\mathcal{B}}_{\tilde{\psi}}(\phi). \end{aligned} \tag{8}$$

Proof Defining $H_q(p) = h(p) - \langle p, h'(q) \rangle$, $\triangle_h(p, q) = H_q(p) - h^*(\phi_q)$. The convex conjugate of H_q is $H_q^*(\phi) = \sup_p \langle p, \phi + h'(q) \rangle - h(p) = h^*(\phi - \phi_q)$, since $h'(q) = \phi_q$. Since q is constant, we get the equality (up to a constant) by plugging H_q^* into Lemma 6. ■

As in Csiszár's divergence, the KL divergence becomes a special case of Bregman divergence by defining h as $h(p) := \int_{\mathcal{T}} p(t) \ln(p(t)) dt$. Thus, we can achieve the same results in Section 3.2 using Bregman divergences as well. Also, it has been shown in various studies that Boosting which minimizes exponential loss can be cast as a special case of Bregman divergence problem with linear equality constraints [8, 11]. An immediate result of Corollary 15, then, is to generalize these approaches by relaxing the equality constraints wrt. various norms and achieve regularized exp-loss optimization problems leading to various regularized boosting approaches. Due to space limitations, we omit the details.

4 Bounds on the Dual Problem and Uniform Stability

Generalization performances of estimators achieved by optimizing various convex functions in Reproducing Kernel Hilbert Spaces have been studied extensively. See e.g. [19, 5] and references therein. Producing similar results in the general form of convex analysis allows us to unify previous results via simpler proofs and tight bounds.

4.1 Concentration of empirical means

One of the key tools in the analysis of divergence estimates is the fact that deviations of the random variable $\tilde{\psi} = \frac{1}{m}\psi(t_i)$ are well controlled.

Theorem 15. *Denote by* $T := \{t_1, \ldots, t_m\} \subseteq \mathcal{T}$ *a set of random variables drawn from* p. *Let* $\psi : \mathcal{T} \to \mathcal{B}$ *be a feature map into a Banach space* $\mathcal{B}$ *which is uniformly bounded by* R. *Then the following bound holds*

$$\left\| \frac{1}{m} \sum_{i=1}^{m} \psi(t_i) - \mathbf{E}_p \left[\psi(t)\right] \right\|_{\mathcal{B}} \leq 2R_m(\mathcal{F}, p) + \epsilon \tag{9}$$

with probability at least $1 - \exp\left(-\frac{\epsilon^2 m}{R^2}\right)$. *Here* $R_m(\mathcal{F}, p)$ *denotes the Rademacher average wrt the function class* $\mathcal{F} := \{\phi_p(\cdot) = \langle \psi(t), \phi_p \rangle \;\; \text{where} \;\; \|\phi\|_{\mathcal{B}^*} \leq 1\}$.

Moreover, if $\mathcal{B}$ *is a RKHS with kernel* $k(t, t')$ *the RHS of (9) can be tightened to* $\sqrt{m^{-1}\mathbf{E}_p\left[k(t,t) - k(t,t')\right]} + \epsilon$. *The same bound for* ϵ *as above applies.*

See [2] for more details and [20] for earlier results on Hilbert Spaces.

Proof The first claim follows immediately from [2, Theorem 9 and 10]. The second part is due to an improved calculation of the expected value of the LHS of (9). We have by convexity

$$\begin{aligned} \mathbf{E}_p \left[\left\| \frac{1}{m} \sum_{i=1}^{m} \psi(t_i) - \mathbf{E}_p \left[\psi(t)\right] \right\|_{\mathcal{B}} \right] &\leq \mathbf{E}_p \left[\left\| \frac{1}{m} \sum_{i=1}^{m} \psi(t_i) - \mathbf{E}_p \left[\psi(t)\right] \right\|_{\mathcal{B}}^2 \right]^{\frac{1}{2}} \\ = \quad m^{-\frac{1}{2}} \sqrt{\mathbf{E}_p \left[\left\| \psi(t) - \mathbf{E}_p \left[\psi(t)\right] \right\|^2 \right]} &= m^{-\frac{1}{2}} \sqrt{\mathbf{E}_p \left[k(t,t) - k(t,t')\right]} \end{aligned}$$

The concentration inequality for bounding large deviations remains unchanged wrt. the Banach space case, where the same tail bound holds. ■

The usefulness of Theorem 16 arises from the fact that it allows us to determine ϵ in the inverse problem. If m is small, it is sensible to choose a large value of ϵ and with increasing m our precision should improve with $O(\frac{1}{\sqrt{m}})$. This gives us a *principled* way of determining ϵ based on statistical principles.

4.2 Stability with respect to changes in b

Next we study the stability of constrained optimization problems when changing the empirical mean parameter b. Consider the convex dual problem of Lemma 6 and the objective function of its special case (7). Both problems can be summarized as

$$L(\phi, b) := f(A\phi) - \langle b, \phi \rangle + \epsilon \left\|\phi\right\|_{\mathcal{B}^*}^k \tag{10}$$

where $\epsilon > 0$ and $f(A\phi)$ is a convex function. We first show that for any b', the difference between the value of $L(\phi, b')$ obtained by minimizing $L(\phi, b)$ with respect to ϕ and vice versa is bounded.

Theorem 16. *Denote by ϕ, ϕ' the minimizers of $L(\cdot, b)$ and $L(\cdot, b')$ respectively. Then the following chain of inequalities holds:*

$$L(\phi, b') - L(\phi', b') \leq \langle b' - b, \phi' - \phi \rangle \leq \|b' - b\|_{\mathcal{B}} \|\phi' - \phi\|_{\mathcal{B}^*} \tag{11}$$

$$\text{and } L(\phi, b) - L(\phi', b') \leq \langle \phi, b' - b \rangle \leq \|b' - b\|_{\mathcal{B}} \|\phi'\|_{\mathcal{B}^*} \tag{12}$$

Proof To show (11) we only need to prove the first inequality. The second one follows by Hölder's theorem:

$$\begin{aligned} L(\phi, b') - L(\phi', b') =& L(\phi, b') - L(\phi, b) + L(\phi, b) - L(\phi', b) + L(\phi', b) - L(\phi', b') \\ \leq& \langle b - b', \phi \rangle + \langle \phi', b' - b \rangle \end{aligned}$$

We used the fact that by construction $L(\phi', b) \geq L(\phi, b)$. To show (12) we use almost the same chain of inequalities, bar the first two terms. ■

In general, $\|\phi - \phi'\|$ can be bounded using Lemma 7,

$$\|\phi' - \phi\|_{\mathcal{B}^*} \leq \|\phi\|_{\mathcal{B}^*} + \|\phi'\|_{\mathcal{B}^*} \leq 2 \left(C/k\epsilon\right)^{\frac{1}{k-1}} . \tag{13}$$

For the special case of $\mathcal{B}$ being a RKHS, however, one can obtain considerably tighter bounds directly on $\|\phi' - \phi\|$ in terms of the deviations in b' and b:

Lemma 17. *Assume that $\mathcal{B}$ is a Hilbert space and let $k = 2, \epsilon > 0$ in (10). Let ϕ and ϕ' be the minimizers of $L(\cdot, b)$ and $L(\cdot, b')$ respectively, where L is defined as in (10). The the following bound holds:*

$$\|\phi - \phi'\| \leq \tfrac{1}{\epsilon} \|b - b'\| \tag{14}$$

Proof The proof idea is similar to that of [6, 19]. We construct an auxiliary function $R : \mathcal{B} \to \mathbb{R}$ via

$$R(z) = \langle A^*[f'(A\phi) - f'(A\phi')] + b' - b, z - \phi' \rangle + \epsilon \|z - \phi'\|^2 .$$

Clearly $R(\phi') = 0$ and R is a *convex* function in z. Taking derivatives of $R(z)$ one can check that its minimum is attained at ϕ:

$$\partial_z R(z) = A^* f'(A\phi) - b - A^* f'(A\phi') + b' + 2\epsilon(z - \phi')$$

For $z = \phi$, this equals $\partial_\phi L(\phi, b) - \partial_{\phi'} L(\phi', b')$ which vanishes due to optimality in L. From this, we have

$$\begin{aligned} 0 &\geq \langle A^*[f'(A\phi) - f'(A\phi')] + b' - b, \phi - \phi' \rangle + \epsilon \|\phi - \phi'\|^2 . \\ &\geq \langle b' - b, \phi - \phi' \rangle + \epsilon \|\phi - \phi'\|^2 \\ &\geq - \|b - b'\| \|\phi - \phi'\| + \epsilon \|\phi - \phi'\|^2 \end{aligned}$$

Here the first inequality follows from $R(\phi') > R(\phi)$, the second follows from the fact that for convex functions $\langle g'(a) - g'(b), a - b \rangle \geq 0$, and the third inequality is an application of Cauchy-Schwartz. Solving for $\|\phi - \phi'\|$ proves the claim. ■

4.3 Risk bounds

We are now in a position to combine concentration and stability results derived in the previous two sections into risk bounds for the values of divergences.

Theorem 18. *Assume that* $b = \frac{1}{m}\sum_{i=1}^{m}\psi(t)$ *and let* $b^* := \mathbf{E}_p[\psi(t)]$. *Moreover, denote by* ϕ, ϕ^* *the minimizers of* $L(\cdot, b)$ *and* $L(\cdot, b^*)$ *respectively. Finally assume that* $\|\psi(t)\| \leq R$ *for all* $t \in \mathcal{T}$. *Then*

$$\|\phi\|\left[2R_m(\mathcal{F}, p) + \epsilon\right] \leq L(\phi^*, b^*) - L(\phi, b) \leq \|\phi^*\|\left[2R_m(\mathcal{F}, p) + \epsilon\right] \tag{15}$$

where each inequality holds with probability $1 - \exp\left(-\frac{\epsilon^2 m}{R^2}\right)$.

Proof Combination of Theorem 16 and (12) of Theorem 17. ■

Note that this is considerably stronger than a corresponding result of [9], as it applies to arbitrary linear classes and divergences as opposed to ℓ_∞ spaces and Shannon entropy. A stronger version of the above bounds can be obtained easily for RKHSs, where the Rademacher average is replaced by a variance bound.

If we want to bound the performance of estimate x with respect to the actual loss $L(\cdot, b^*)$ rather than $L(\cdot, b)$ we need to invoke (11). In other words, we show that on the true statistics the loss of the estimated parameter cannot be much larger than the loss of true parameter.

Theorem 19. *With the same assumptions as Theorem 19 we have with probability at least* $1 - \exp\left(-\frac{\epsilon^2 m}{R^2}\right)$

$$L(\phi, b^*) - L(\phi^*, b^*) \leq 2\left(\frac{C}{k\epsilon}\right)^{\frac{1}{k-1}}\left(2\mathcal{R}_n(\mathcal{F}_{\mathcal{B}}) + \epsilon\right). \tag{16}$$

Here C *is the Lipschitz constant of* $f(A\cdot)$. *If* $\mathcal{B}$ *is an RKHS we have with probability at least* $1 - \exp\left(-\frac{\epsilon^2 m}{50R^4}\right)$ *for* $m \geq 2$

$$L(\phi, b^*) - L(\phi^*, b^*) \leq \frac{1}{\epsilon}\left[\frac{1}{m}\mathbf{E}_p\left[k(t,t) - k(t,t')\right] + \epsilon\right]. \tag{17}$$

Proof To prove (16) we use (11) which bounds

$$L(\phi, b^*) - L(\phi^*, b^*) \leq \|b^* - b\|_{\mathcal{B}}\left(\|\phi\|_{\mathcal{B}^*} + \|\phi^*\|_{\mathcal{B}^*}\right).$$

The first factor is bounded by (9) of Theorem 16. The second term is bounded via Lemma 7. A much tighter bound is available for RKHS. Using (11) in conjunction with (14) of Lemma (18) yields

$$L(\phi, b^*) - L(\phi^*, b^*) \leq \frac{1}{\epsilon}\|b - b^*\|^2$$

We establish a bound for $\|b - b^*\|^2$ by a standard approach, i.e. by computing the mean and then bounding the tail of the random variable. By construction

$$\mathbf{E}\left[\|b - b^*\|^2\right] = \mathbf{E}\left[\left\|\frac{1}{m}\sum_{i=1}^{m}\psi(t_i) - \mathbf{E}\left[\psi(t)\right]\right\|^2\right] = \frac{1}{m}\mathbf{E}\left[\|\psi(t) - \mathbf{E}\left[\psi(t')\right]\|^2\right]$$

Using $k(t,t') = \langle\psi(t), \psi(t')\rangle$ yields the mean term. To bound the tail we use McDiarmid's bound. For this, we need to check by how much $\|b - b^*\|^2$ changes if we replace one term $\psi(t_i)$ by an arbitrary $\psi(t_i')$ for some $t_i' \in \mathcal{T}$. We have

$$\begin{aligned}&\left\|b + \tfrac{1}{m}(\psi(t_i') - \psi(t_i)) - b^*\right\|^2 - \|b - b^*\|^2\\ \leq &\tfrac{1}{m}\left\|\psi(t_i') - \psi(t_i)\right\|\left\|2(b + b^*) + \tfrac{1}{m}(\psi(t_i') - \psi(t_i))\right\| \leq 10R^2/m\end{aligned}$$

for $m \geq 2$. Plugging this into McDiarmid's bound yields that $\|b - b^*\|^2$ deviates from its expectation by more than ϵ with probability less than $\exp\left(-\frac{m\epsilon^2}{50R^4}\right)$. ■

Theorem 20 also holds for $\mathfrak{L}_\psi^B$. Since the KL divergence is an example of Csiszár's divergence, using this bound allows us to achieve stability results for MAP estimates immediately.

5 Optimization Algorithm and Convergence Properties

In the most general form, our primal problem, $f(x)$ subject to $\|Ax - b\|_{\mathcal{B}} \leq \epsilon$ is an abstract program, where both the constraint space $\mathcal{B}$ and the domain $\mathcal{X}$ may be infinite, i.e. both the primal and the dual turn out to be infinite problems. Thus, except for special cases finding an optimal solution in polynomial time may be impossible. It turns out that a sparse greedy approximation algorithm proposed by Zhang [23] is an efficient way of solving this class of problems efficiently, providing good rates of convergence (in contrast, the question of a convergence *rate* remains open in [9]).

Algorithm 1 Sequential greedy approximation [23]

1: **input:** sample of size n, statistics b, base function class $\mathcal{B}_{\text{base}}^*$, approximation ϵ, number of iterations K, and radius of the space of solutions R
2: Set $\phi = 0$.
3: **for** $k = 1, \ldots, K$ **do**
4: Find $(\hat{\imath}, \hat{\lambda})$ such that for $e_i \in \mathcal{B}_{\text{base}}^*$ and $\lambda \in [0,1]$ the following is approximately minimized:
$$L((1-\lambda)\phi + \lambda R e_i, b)$$
5: Update $\phi \leftarrow (1 - \hat{\lambda})\phi + \hat{\lambda} R e_{\hat{\imath}}$
6: **end for**

Algorithm 1 requires that we have an efficient way of updating ϕ by drawing from a base class of parameters $\mathcal{B}_{\text{base}}^*$ which "generates" the space of parameters

$\mathcal{B}^*$. In other words, we require that $\text{span}\mathcal{B}^*_{\text{base}} = \mathcal{B}^*$. For instance we could pick $\mathcal{B}^*_{\text{base}}$ to be the set of vertices of the unit ball in $\mathcal{B}^*$.

Note that Step 4 in Algorithm 1 only needs to be approximate. In other words, we only need to find $(\hat{\imath}, \hat{\lambda})$ such that the so-found solution is within δ_k of the optimal solution, as long as $\delta_k \to 0$ for $k \to \infty$.

Also note the dependency on R: one needs to modify the setting of [23] to make it applicable to arbitrary convex sets. As long as R is chosen sufficiently large such as to include the optimal solution the conditions of [23] apply.

Theorem 20 ([23, Theorem II.1]). *Let M_β be an upper bound on $L''(\phi)$. If the optimization is performed exactly at each step (i.e. $\delta_k = 0$ for all k) we have*

$$L(\phi^k, b) - L(\hat{\phi}, b) \leq 2M/(k+2) \tag{18}$$

where $\hat{\phi}$ is the true minimizer of $L(\phi, b)$.

This has an interesting implication when considering the fact that deviations between the optimal solution of $L(\phi^*, b^*)$ for the true parameter b^* and the solution achieved via $L(\phi, b)$ are $O(1/\sqrt{m})$, as follows from Section 4.3. It is essentially pointless to find a better solution than within $O(1/\sqrt{m})$ for a sample of size m. Hence we have the following corollary:

Corollary 21. *Zhang's algorithm only needs $O(\sqrt{m})$ steps for a set of observations of size m to obtain almost optimal performance.*

When the dual is a finite program, it is possible to achieve linear convergence rates (where the difference in Equation 18 goes exponentially fast to 0 in k) [17]. The obvious special case when the dual is a finite dimensional optimization problem is when the index set I over the statistics is finite.

Consider $\mathcal{X}$ itself is a finite dimensional problem, for example, when we want to estimate the conditional density $p(y|x)$ of a classification task wrt. inequality constraints in a Banach space. In that case, our primal is a semi-infinite program (SIP), i.e. optimization over a finite dimensional vector space wrt infinite number of constraints. Then, using a Helly-type theorem, one can show that the SIP can be reduced to a finite program (i.e. with finite number of constraints) and we immediately get a finite dual program. This is a generalization of a family of results commonly referred to as representer theorems.

6 Conclusion

Our generalized framework of convex duality allowed us to unify a large class of existing inference algorithms in a common framework, to provide statistical bounds for the estimates, and to provide a practical algorithm.

Note that in the present paper we barely scratched the surface of alternative divergence measures, such as Tsallis or Sharma-Mittal entropy. Also, we did not discuss in detail what becomes of structured estimation methods when applied in conjunction with Zhang's algorithm. Likewise, the connection between Boosting

and an approximate solution of inverse problems has not been explored yet. Finally, it may be possible to minimize the divergence directly in transductive settings. We expect this set of problems to be a fertile ground for future research.

Acknowlegements: We thank Tim Sears, Thomas Gaertner and Vishy Vishwanathan. National ICT Australia is funded through the Australian Government's *Baking Australia's Ability* initiative, in part through the Australian Research Council. This work was supported by the PASCAL Network of Excellence.

References

1. Y. Altun, T. Hofmann, and A. J. Smola. Exponential families for conditional random fields. In *Uncertainty in Artificial Intelligence UAI*, pages 2–9, 2004.
2. K. Borgwardt, A. Gretton, and A.J. Smola. Kernel discrepancy estimation. In *Annual Conference on Computational Learning Theory COLT*, 2006. submitted.
3. J. Borwein and Q.J. Zhu. *Techniques of Variational Analysis*. Springer, 2005.
4. B.E. Boser, I.M. Guyon, and V.N. Vapnik. A training algorithm for optimal margin classifiers. In *COLT'92*, pages 144–152, 1992.
5. O. Bousquet, S. Boucheron, and G. Lugosi. Theory of classification: a survey of recent advances. *ESAIM: Probability and Statistics*, 2004. submitted.
6. O. Bousquet and A. Elisseeff. Stability and generalization. *JMLR*, 2:499–526, 2002.
7. S. Chen and R. Rosenfeld. A Gaussian prior for smoothing maximum entropy models. Technical Report CMUCS-99-108, Carnegie Mellon University, 1999.
8. M. Collins, R. E. Schapire, and Y. Singer. Logistic regression, adaboost and bregman distances. In *COLT'00*, pages 158–169, 2000.
9. M. Dudik, S. Phillips, and R. E. Schapire. Performance guarantees for regularized maximum entropy density estimation. In *Proceedings of COLT'04*, 2004.
10. M. P. Friedlander and M. R. Gupta. On minimizing distortion and relative entropy. *IEEE Transactions on Information Theory*, 52(1), 2006.
11. J. Kivinen and M. Warmuth. Boosting as entropy projection. *COLT* 1999
12. J. Lafferty. Additive models, boosting, and inference for generalized divergences. In *COLT '99*, pages 125–133, New York, NY, USA, 1999. ACM Press.
13. Q. V. Le, A. J. Smola, and S. Canu. Heteroscedastic gaussian process regression. In *International Conference on Machine Learning ICML 2005*, 2005.
14. V.A. Morozov. *Methods for solving incorrectly posed problems*. Springer, 1984.
15. R. Neal. Priors for infinite networks. Tech. Rep. CRG-TR-94-1, U. Toronto, 1994.
16. I. Nemenman and W. Bialek. Occam factors and model independent bayesian learning of continuous distributions. *Physical Review E*, 65(2):6137, 2002.
17. G. Rätsch, S. Mika, and M.K. Warmuth. On the convergence of leveraging. In *Advances in Neural Information Processing Systems (NIPS)*, 2002.
18. D. L. Ruderman and W. Bialek. Statistics of natural images: Scaling in the woods. *Phys Rev. Letters*, 1994.
19. B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, 2002.
20. J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.
21. A. N. Thikhonov and V. Y. Arsenin. *Solutions of Ill-Posed Problems*. Wiley, 1977.
22. M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. Technical Report 649, UC Berkeley, September 2003.
23. T. Zhang. Sequential greedy approximation for certain convex optimization problems. *IEEE Transactions on Information Theory*, 49(3):682–691, March 2003.