Mirror Descent Introduction

0. References

for "math people": see the Nemirovski & Yudin 1983 book and/or their recent papers

for people "tired reading proofs": see the online learning survey by Shalev-Shwartz (2012)
— does have proofs but much less intense!

1. From Gradient Descent to Mirror Descent

Goal: minimise some loss function

$$\mathcal{L}(\lambda; D) = \mathbb{E}_{x \sim D}[l(\lambda; x)]$$

given the dataset D and loss measure $l(\lambda; x)$, λ is the parameter of the model.

① gradient descent:

$$\lambda_{t+1} \leftarrow \lambda_t - \beta_t \nabla \mathcal{L}(\lambda_t)$$

some learning rate at time t

note: can be sub-gradient $\partial \mathcal{L}$

② an equivalent optimization problem (unconstrained)

$$\lambda_{t+1} = \arg\min_{\lambda} \Big\{ \underbrace{\langle \lambda, \nabla \mathcal{L}(\lambda_t) \rangle + \frac{1}{2\beta_t} \|\lambda - \lambda_t\|_2^2}_{\triangleq \hat{\mathcal{L}}(\lambda; \lambda_t)} \Big\}$$

to see this: set

$$0 = \nabla \hat{\mathcal{L}}(\lambda) = \nabla \mathcal{L}(\lambda_t) + \frac{1}{\beta_t}(\lambda - \lambda_t)$$

$$\Rightarrow \lambda \leftarrow \lambda_t - \beta_t \nabla \mathcal{L}(\lambda_t)$$

③ extending ②: mirror descent (MD)

we change the L_2 measure in $\hat{\mathcal{L}}$ to some other divergence!

In particular we're interested in the Bregman divergence

$$B_\psi(\lambda, \lambda') = \psi(\lambda) - \psi(\lambda') - \langle \lambda - \lambda', \nabla \psi(\lambda') \rangle$$

a (strongly) convex and (twice-)differentiable function

now new problem (MD)

$$\lambda_{t+1} = \arg\min_{\lambda} \Big\{ \langle \lambda, \nabla \mathcal{L}(\lambda_t) \rangle + \frac{1}{\beta_t} B_\psi(\lambda, \lambda_t) \Big\}$$

Mirror Descent Intro. (cont.)

new MD problem:

$$\lambda_{t+1} = \arg\min_{\lambda} \left\{ \langle \lambda, \nabla \mathcal{L}(\lambda_t) \rangle + \frac{1}{\beta_t} B_\psi(\lambda, \lambda_t) \right\}$$

we solve it by zeroing the gradient:

$$0 = \nabla \hat{\mathcal{L}}(\lambda) = \nabla \mathcal{L}(\lambda_t) + \frac{1}{\beta_t}\left[\nabla\psi(\lambda) - \nabla\psi(\lambda_t)\right]$$

$$\Rightarrow \quad \nabla\psi(\lambda_{t+1}) \leftarrow \nabla\psi(\lambda_t) - \beta_t \nabla\mathcal{L}(\lambda_t)$$

④ examples:

1) L_2 measure: set $\psi(\lambda) = \frac{1}{2}\|\lambda\|_2^2$,

easy to verify $B_\psi(\lambda, \lambda') = \frac{1}{2}\|\lambda - \lambda'\|_2^2$

2) KL-divergence for general distributions:

Set $\psi(p) = -H(p) = \int p \log p \, d\mu$,

easy to verify $B_\psi(p, q) = KL[p\|q]$

3) KL-divergence for exponential families' natural parameters:

set $p_\lambda(\theta) = \exp[\langle \lambda, \Phi(\theta) \rangle - A(\lambda)]$

and $\psi(\lambda) = A(\lambda)$,

then $B_\psi(\lambda, \lambda') = KL[p_{\lambda'} \| p_\lambda]$

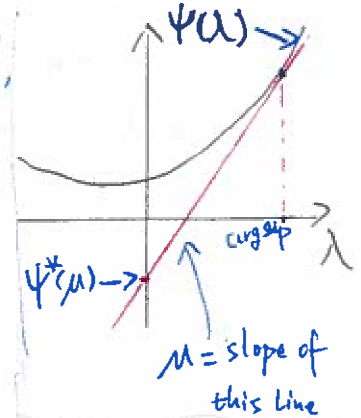
⑤ convex conjugate interpretation

(Fenchel-Legendre transform)

$$\psi^*(\mu) = \sup_{\lambda \in \Lambda} \{ \langle \lambda, \mu \rangle - \psi(\lambda) \},$$

$$\mu \in \Lambda^* \triangleq \mathcal{M}$$

and $\psi^{**}(\lambda) = \psi(\lambda) = \sup_{\mu \in \mathcal{M}} \{ \langle \lambda, \mu \rangle - \psi^*(\mu) \}$.



Importantly, we have:

$$\begin{cases} \nabla\psi(\lambda) = \mu \\ \nabla\psi^*(\mu) = \lambda \end{cases}$$

(example: natural parameter λ, moment parameter μ)

Now we rewrite the MD steps:

$\mu_{t+1} \leftarrow \mu_t - \beta_t \nabla\mathcal{L}(\lambda_t)$ (gradient step)

$\lambda_{t+1} \leftarrow \nabla\psi^*(\mu_{t+1})$ (mirror step)

$\Rightarrow$ do gradient descent in the dual space while the gradients are evaluated in the primal space.

Mirror Descent Intro. (cont.)

⑥ Connection to Natural Gradient Descent (NGD)

(Amari 1998 paper)

- the original gradient descent assumes Euclidean space with local metric tensor $G(\lambda) \equiv I$ the identity matrix
- the steepest descent in a Riemannian manifold with metric tensor $G(\lambda)$:

$$\lambda_{t+1} \leftarrow \lambda_t - \beta_t G^{-1}(\lambda_t) \nabla \mathcal{L}(\lambda_t)$$

$\boxed{\text{Thm 1. MD is NGD on manifold } (\mathcal{M}, \nabla^2 \psi^*).}$

Proof.

$$\nabla_{\mu_t} \mathcal{L}(\lambda_t) = \nabla_{\mu_t} \lambda_t \nabla_{\lambda_t} \mathcal{L}(\lambda_t)$$
$$= \nabla^2 \psi^*(\mu_t) \nabla_{\lambda_t} \mathcal{L}(\lambda_t)$$

so in MD, use $\lambda_t = \nabla \psi^*(\mu_t)$

$$\mu_{t+1} \leftarrow \mu_t - \beta_t \nabla \mathcal{L}(\lambda_t)$$
$$= \mu_t - \beta_t [\nabla^2 \psi^*(\mu_t)]^{-1} \nabla_{\mu_t} \mathcal{L}(\nabla \psi^*(\mu_t))$$

$\square$

2. MD stochastic approximation methods

Recall $\mathcal{L}(\lambda; D) = \mathbb{E}_{x \sim D}[l(\lambda; x)]$,

then just use $\underset{\text{with } x \sim D}{l(\lambda; x)} \approx \mathcal{L}(\lambda; D)$.

MD SA:

$$\lambda_{t+1} = \arg\min_{\lambda} \left\{ \langle \lambda, \nabla l(\lambda_t; x_t) \rangle + \frac{1}{\beta_t} B_\psi(\lambda, \lambda_t) \right\}$$

solution: $\mu_{t+1} \leftarrow \mu_t - \beta_t \nabla l(\lambda_t; x_t)$, $x_t \sim D$

$\lambda_{t+1} \leftarrow \nabla \psi^*(\mu_{t+1})$

existing theory:

check paper by Nemirovski et al.

"Robust Stochastic Approximation Approach to Stochastic Programming".

Requiring $B_\psi(\cdot, \lambda)$ to be α-strongly convex wrt. for every λ some norm/divergence $d(\cdot, \cdot)$